DOCUMENT RESUME

ED 386 916 FL 022 933

AUTHOR Thompson, Irene
TITLE Testing Listening Comprehension.
REPORT NO ISSN-0001-0251
PUB DATE Apr 95
NOTE 10p.; AATSEEL = American Association of Teachers of Slavic and East European Languages.
PUB TYPE Journal Articles (080)
JOURNAL CIT AATSEEL Newsletter; v37 n5 p24-31 Apr 1995

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Advance Organizers; *Cognitive Processes; *Evaluation Criteria; *Language Tests; *Listening Comprehension; Listening Comprehension Tests; Multiple Choice Tests; Second Language Instruction; Second Language Learning; Sentence Structure; *Test Construction; Vocabulary
IDENTIFIERS Pausing (Speech)

ABSTRACT

This article discusses practical considerations in developing tests of listening comprehension in second language learning with a particular emphasis on the choice of listening passages and assessment tasks. The listening construct is defined as the process of receiving, attending to, and assigning meaning to aural stimuli. Questions should be developed while listening to, not reading, the intended passage. Selection of passage may depend upon orality versus literacy, audio versus video, passage length, content familiarity, vocabulary and sentence structure, elaborations and redundancies, speech rate and pauses, and fuzzy word boundaries or other speech phenomena. Consideration must be given to type of response expected, including multiple choice, true-false, open-ended, recall, and nonverbal. Presentation effects should also be part of the decision, such as advance organizers, language of instructions and questions, and uniformity of presentation. New tests should be refined by pilot testing before actual use. (Contains 41 references.) (NAV)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 386 916

Testing Listening Comprehension

By Irene Thompson

FL022933

BEST COPY AVAILABLE

TESTING LISTENING COMPREHENSION

by Irene Thompson

INTRODUCTION

The central role of listening comprehension in second language (L2) acquisition is now largely accepted, and most modern materials and methodologies are placing an increasing emphasis on activities designed to promote the development of this important skill (Rubin, 1994). Listening comprehension testing, on the other hand, continues to remain somewhat of a neglected area. To begin the discussion of testing L2 listening comprehension, we first need to define the construct. For purposes of this discussion, I will adopt a very general definition proposed by Wolvin and Coakley (1985:74) that listening comprehension is "the process of receiving, attending to, and assigning meaning to aural stimuli." I will then discuss some practical considerations in developing tests of listening comprehension with particular emphasis on the choice of listening passages and assessment tasks.

SPECIAL QUALITIES OF THE AURAL MEDIUM

When developing tests of listening comprehension you should consider the special qualities of the aural medium. To begin with, listeners, unlike readers, cannot review and reevaluate information presented to them. They must comprehend the text as they listen to it, retain information in memory, integrate it with what follows, and continually adjust their understanding of what they hear in the light of prior knowledge and of incoming information. This heavy processing load makes listening comprehension different from reading comprehension in a number of significant ways.

First of all, people recall less information from listening than from reading in terms of both quantity and quality. Although the probability of recalling idea units after both listening and reading is influenced by their position in the hierarchical structure of the text, this effect is more pronounced in the case of listening (Hildyard and Olson, 1982; Lund 1991a; Meyer and McConkie, 1973). Facts that are incidental or irrelevant to the main ideas of the text have a low probability of recall in listening (Shohamy and Inbar, 1991).

This has practical implications for testing listening comprehension. You should put yourself in the position of the examinees and develop the questions as you listen to the passage, not as you read the transcript. This will lessen the likelihood of including questions that are better suited for testing reading than testing listening.

SELECTING LISTENING PASSAGES

Among aural passages are conversations, instructions, announcements, stories, lectures, news reports, movies, plays, interviews, debates, speeches, interviews, and advertisements to mention just a few. Each of these texts has its own special features which affect ways in which it will be processed and understood.

There are many considerations in selecting suitable passages for testing listening comprehension. The most obvious ones are level of difficulty, interest, and relevance. Finding an authentic passage at the desired level of difficulty is not easy because so many factors need to be considered. Bear in mind that it is often impossible to predict the empirical difficulty of listening items on the basis of passages alone, because difficulty resides not just in the text, but in the interaction of text variables with tasks, background knowledge, memory, and inferencing ability. As a result, the same passage can yield items with different degrees of difficulty. Some of the features to keep in mind when selecting listening passages for testing are discussed below.

Orality vs. literacy

Oral texts can be arranged along a continuum with those closer to the spoken language, at one end, and those closer to the written language, at the other (Tannen, 1982, 1985). Idea units in the spoken language are typically expressed in short clauses, are loosely strung together, contain repetitions, and are bounded by pauses because speakers don't always have time to plan their utterances. Idea units in the written language, on the other hand, tend to be longer, more complex, and contain densely packed information because writers have time for planning, editing, and revising (Chafe, 1985). It has been demonstrated that texts closer to the oral end of the continuum yield higher scores on listening comprehension tests than passages closer to the written end. Shohamy and Inbar (1991) showed that with the topic held constant, news broadcasts (pre-written edited monologues) were more difficult to understand than lectures (monologues delivered from written notes). Thompson (1993) reported that conversations yielded higher comprehension scores than expository passages on the listening portion of the *ETS Comprehensive Russian Proficiency Test* (1990). On the other hand, Berne (1992) found no significant difference between scores on a long lecture and an interview on the same topic, both created from the same written article. Other research shows that texts are easier to understand if they contain such conversational features are repeated nouns (Chaudron, 1983), and advance organizers that call attention to major propositions, transitions, and emphases in the text (Chaudron and Richards, 1986). Other spoken features, such as redundancies and elaborations, are helpful only after learners have reached a certain level of proficiency (Chiang and Dunkel, 1992; Derwing, 1989).

If you are planning to use authentic passages for lower-ability examinees, you should look for texts that are closer to the spoken than to the written language. In general, you should avoid using written materials for testing listening comprehension since it is quite difficult to modify them to make them resemble spoken language. Rather than collecting written sources, you should keep a library of recorded passages from radio, TV, movies, or other sources.

Audio or video?

If you decide to base your listening comprehension test on a video segment, you should consider the extent to which visual clues interact with the oral message (Joiner, 1990; Phillips, 1990). Keep in mind that visual support is particularly helpful for lower- proficiency listeners (Mueller, 1980). Videos vary in the extent to which they provide visual support that is helpful to viewers. At the one extreme are segments in which visuals obviate the need for listening, while at the other extreme are segments in which the visuals bear no relationship to the sound track. The extent of visual support varies according to genre, with dramatic segments, such as movies, soap operas, and TV series, providing more visual, action and interaction cues than interviews, speeches, and news, which tend to be dominated by "talking heads." Weather, sports, and various news reports vary in the amount of visual support from segment to segment, and country to country. High-tech American and European TV programs, which abound in location shots, are generally richer in visual cues than programs from Russia, the former republics, and Eastern Europe.

Length of passages

Heavy processing requirements imposed by the oral medium cause listeners to lose concentration rather quickly. Listeners report "tuning out" if passages are more than two-three minutes long (Thompson and Rubin, forthcoming). During the field testing of the *ETS Advanced Russian Listening/Reading Test* (1986) which contained a 50-minute listening and a 50-minute reading portion, students not only did more poorly in listening than in reading, but they also reported greater difficulty maintaining their concentration during 50 minutes of listening than during an equivalent period of reading (unpublished data).

Experience shows that listeners can attend to some types of oral passages longer than to others. For instance, dramatic TV segments, which consist of conversations accompanied by action, hold listeners' attention longer than TV news reports, speeches, or lectures. As a rule of thumb, oral passages for testing should not be longer than two or three minutes.

Content familiarity

The content of a listening passage will affect all test takers by making it easier to understand for those who are familiar with the topic, and more difficult for those who are not (Chiang and Dunkel, 1992; Long, 1990; Markham and Latham, 1987; Schmidt-Rinehard, 1992). This is especially true if test questions require students to go beyond the passage, and to make inferences based on prior knowledge about the subject (Buck, 1991). To minimize the effect of prior knowledge on listening test performance, you should either select passages that are neutral with respect to potential differences in familiarity with the topic, or to include an extensive sampling of topics.

Vocabulary

There is little doubt that vocabulary recognition plays an extremely important role in listening comprehension. Passages which contain frequently used words are easier to understand than passages which contain many specialized and technical words, idioms, and cultural allusions. Being able to recognize a familiar word which has little to do with the main idea of the passage can cause lower-level listeners to "go off on a tangent," as illustrated in the following example. First-year students of Russian listened to a conversation between two Muscovites making plans to attend a friend's birthday party. Among other details, they agreed to meet at the "Tretyakovky" metro station. When asked "**What is this conversation about?**", some students answered that it was about going to a museum, because they recognized the word **Tretyakovsky**, the name of a famous art gallery.

When selecting listening passages for lower-proficiency test-takers, you should make sure that some of the key vocabulary is recognizable, or inferable from context. Keep in mind, however, that familiar words and cognates are not always easily retrievable from dynamic speech, and that even fairly advanced learners may fail to understand familiar words if the latter are used in a different meaning or in an unfamiliar context, and may experience difficulties with numbers and proper names (Laviosa, 1991).

Sentence structure

A question test constructors often ask is "Should I simplify sentence structures to make the passage easier to comprehend?" It seems intuitively appealing to think that syntax should play a major role in listening comprehension, but there is not enough research to answer the question as to whether everything else being equal, syntactically complex sentences are harder to understand than simple ones. Blau (1990) found no significant effect of sentence structure simplification on listening comprehension of advanced ESL students, while Glisan (1985) found that longer, modified sentences were actually better understood than shorter, unmodified ones by advanced students of Spanish. Unfortunately, there are no studies that deal with the effects of syntactic complexity on the listening comprehension of lower-ability L2 listeners.

There is some evidence, however, that word order may affect the comprehension of speech. For instance, advanced

English-speaking students of Spanish comprehended Spanish Subject-Verb-Object (SVO) sentences better than VSO and OVS sentences (Glisan, 1985). The latter type was particularly difficult, leading one to hypothesize that passages in which there are many OVS sentences (such as is often the case in Russian), might be difficult to process for speakers of English where this pattern is extremely uncommon.

Elaborations and redundancies

Redundancy in the form of repeated nouns ("**The pencil... the pencil is on the table**") appears to be more effective than other reinstatement devices, such as synonyms or simple topic reiterations ("**This is a pencil. The pencil is on the table**") for listeners at lower and intermediate levels of proficiency (Chaudron, 1983). Increased redundancy of information (repetition) and elaboration (paraphrase, use of synonyms) may not be beneficial for lower-ability listeners because lack of adequate vocabulary prevents them from taking advantage of redundant information (Chiang and Dunkel, 1992). The practical implication is that an authentic passage can be made more comprehensible for lower-proficiency learners through added repetition of nouns, while for more advanced listeners paraphrase and modifiers may be more effective.

Insertion of various macro discourse markers referring to major propositions in a monologue may also improve its comprehensibility. Examples of macro discourse markers are "**What I'm going to talk about today is....**," or "**Let's go back to the beginning.**" On the other hand, micro discourse markers, such as temporal links (**after that**) and causal connectors (**therefore, consequently**) signaling intersentential connections may have no facilitating effect (Chaudron and Richards, 1986; Hron et al., 1985). The practical implication is that a passage can be made more accessible if insert macro markers are inserted at major discourse boundaries.

Speech rate

There is some rather unsurprising evidence that excessive speed (faster than 200 wpm) impairs comprehension of lower-intermediate ESL learners (Griffith, 1990). These learners seem to perform best at a slower rate of around 120 wpm (Griffith, 1992; Kelch, 1985). On the other hand, more advanced listeners appear to be affected not so much by rate of speech as by other factors, such as text type, task, and prior knowledge (Blau, 1990; King and Behnke, 1989). Keep in mind that research evidence is limited and conflicting because studies use different subjects, languages, texts, tasks, definitions of "normal" rate for different languages, and measurement techniques. However, it seems reasonable to assume that passages delivered at high speech rates are, probably, not suitable for examinees at lower levels of proficiency.

Pauses

Since spoken language tends to be relatively seamless and continuous, pauses act much like punctuation marks do in writing to break up the spoken signal into constituents. Therefore, one would assume that pauses should help listeners process the message more easily. However, studies indicate that there appears to be a threshold of language proficiency below which pauses do not aid listening comprehension. For instance, pause insertion did not increase the comprehension of lower-ability students (Jacobs et al., 1988), but inserting longer than normal pauses at clause or sentence boundaries helped advanced listeners to comprehend expository passages more than slowing down the speech rate (Blau, 1990, 1991).

Fuzzy word boundaries and other dynamic speech phenomena

Words in dynamic speech undergo various transformations through assimilation, vowel reduction, consonant weakening, liaison, and syllable contraction, so that even native listeners have occasional difficulty in reconstructing citation forms from a stream of speech (Hieke, 1987). In addition, units in dynamic speech, i.e., uninterrupted stretches of speech between pauses, are much longer than citation forms, i.e., units corresponding to single words. According to Carterette and Jones (1974:367), dynamic forms contain an average of twelve phonemes, as compared to citation forms that contain an average of just three. L2 listeners whose initial exposure is often to L2 words spoken in isolation, fail to recognize even highly familiar words in running speech because their limited knowledge of the language does not allow them to compensate for missing phonological information due to assimilation, contraction, liaison, and elision (Henrichsen, 1990). In Russian, words can change both in terms of the number of syllables and in vowel and consonant quality. Thus, **[stól]** can be buried in **[nəstʌl'é]**. This is one more reason why one should not depend on written transcripts when selecting listening passages. One should listen, instead, to the spoken version to decide whether the passage contains too many phonological transformations to be suitable for lower-proficiency learners. You may need to re-record a passage in which key vocabulary items have undergone such significant sandhi-transformations as to be inaccessible to lower-level listeners.

DESIGNING ASSESSMENT TASKS

If you want to interpret scores on tests of listening comprehension as indicators of listening ability, you must make sure that these scores measure listening ability and not much else. This means that you should minimize potential sources of measurement error, i.e., factors other than listening comprehension. Various sources of measurement error in testing listening comprehension are discussed below.

Memory

Memory is an inseparable part of comprehension. However, its role in listening may be different from its role in reading. In reading, the examinee can refer back to portions of the text that contain information necessary for answering a question. In listening, however, the examinee cannot reaccess the text when attempting to construct an answer. This means that you should consider the extent to which a question may overburden the examinee's ability to remember textual information (Thompson, 1993). A listener may have comprehended what was being said at the time of listening, but by the time he or she got to the question(s), the memory trace may have been erased by subsequent information in the text, and by having to read the question and answer options. In real life, note-taking is of considerable help to listeners, but under the time constraints of a testing situation, careful note-taking may not always be possible. An example from an experimental Russian listening comprehension test taken by 100 students (unpublished data) shows why two questions based on the same passage have different difficulty levels due to differential memory load. After having listened to a weather report, students were asked two multiple-choice questions which are reproduced below:

1. The forecast calls for
   (A) sunshine
   (B) light snow
   (C) partial overcast
   (D) thick fog

2. The current temperature in Moscow is
   (A) 6 degrees
   (B) 10 degrees
   (C) 13 degrees
   (D) 19 degrees

Ninety-six percent of the examinees answered the first question correctly, in contrast to the second question which was answered correctly by seventy-eight percent of the test-takers. Why was the second question more difficult than the first one? One possibility is that the answer to the first question depended largely on being able to recognize a specific vocabulary item, while the response to the second question required the examinees to recall which number corresponded to the current temperature, as opposed to barometric pressure, wind velocity, and nighttime temperature, all of which were also mentioned in the forecast. This means that you should make an effort to design items that do not require listeners to recall incidental details (Aly, 1993).

Inferencing and other mental operations

Test questions measure not only comprehension but also the ability to draw inferences, solve problems, and make deductions from text content. An example from a Russian test shows how cognitive demands can affect item difficulty. After listening to a monologue about Pasternak's novel *Doctor Zhivago*, examinees were asked three questions which test developers predicted to be roughly equivalent in difficulty. The results of the field test proved them wrong. Two of the questions which dealt with information that was explicitly stated in the monologue were answered correctly by about half of the test takers. However, only ten percent of them were able to answer the third question which required them to make an inference. This suggests that test developers should keep in mind that the more complex the mental operations involved in arriving at the correct answer, the more difficult the listening item is likely to be.

TYPE OF EXPECTED RESPONSE

Listeners' performance will be affected by the type of response that is required of them. Among the most commonly used responses are selected responses and constructed responses. Selected responses do not require test-takers to create a response, merely to select an the most plausible option. Constructed responses require test-takers to produce their own answers. Berne (1992) found that students of Spanish received significantly higher scores on a multiple-choice version than on either an open-ended or cloze versions of the same test, but no difference between open-ended and cloze versions. In a validation study of the ACTFL Russian proficiency guidelines, the mean score for multiple-choice questions was higher than that for open-ended items.

The advantages and disadvantages associated with different types of responses are discussed below.

Multiple-choice questions

Multiple-choice questions have several advantages. In the first place, they are easy and fast to score because no judgment is required on the part of the scorers. Secondly, multiple-choice items require a minimal amount of time to complete, therefore, multiple-choice tests can include many items, which enhances test reliability. Thirdly, multiple-choice items minimize the confounding of listening with speaking or writing because they have no production requirements, even though reading remains a confounding factor. All these features make multiple-choice tests practical in situations that require testing of large numbers of individuals. However, there are a number of disadvantages as well. First, multiple-choice items invite guessing. Secondly, important parts of a passage sometimes cannot be tested simply because three plausible distractors cannot be found. Last, but not least, good multiple-choice questions are extremely difficult to write. Common problems include clues pointing to the right answer, confusing or implausible distractors, insufficient number of distractors—ideally, there should be one correct answer and three distractors—, unclear or lengthy wording, negative wording, and more than one correct option.

True-false questions

True-false items are easier to write than multiple-choice questions, but the examinee has a fifty-percent chance of being correct by guessing. Because both multiple-choice and true-false responses encourage guessing, it is common practice for test instructions to state whether or not there is a penalty for guessing, and what that penalty is.

Open-ended questions

Open-ended questions avoid some of the problems associated with multiple-choice items. In the first place, they invite guessing less than multiple-choice items. In the second place, they allow test constructors to ask any question, not just a question for which four plausible multiple-choice options can be designed. However, open-ended questions do not always work as intended because more than one answer can sometimes be reasonably interpreted as correct. This often happens when the answer depends on extratextual information— a situation which frequently arises in connection with higher-level questions. Since test-takers differ in terms of background knowledge, it is sometimes difficult to predict what their answers might be. Here is an example of a poorly designed open-ended question. After listening to an interview with a literary critic, test-takers were asked: "**What is Solzhenitsyn's role as a writer?**" Some test-takers based their answers on prior knowledge about Solzhenitsyn and not on what was actually stated in the interview. As a result, it was difficult to decide whether some answers were acceptable or not. To solve this problem, the question was re-worded to read: "**What arguments did the interviewee use to support her opinion about Solzhenitsyn's writing?**" This formulation indicated to the test-takers that their answer had to be based on information contained in the interview. As a result, the range of responses was narrowed, and scoring was made easier.

Another problem with open-ended questions arises when there is insufficient indication of just how much information should be included in the answer (Buck, 1991). Here is an example. Students listened to a monologue in which the speaker outlined a program for economic reforms in Russia. They were asked: "**How does the speaker propose to change Russia's economy?**" Answers ranged from skeletal ("**He advocates capitalism**") to relatively detailed ("**He suggests that state enterprises be converted to private ownership; he also wants the government to attract foreign investments and to control inflation**"). Binary (right/wrong) scoring would have been inappropriate in this case because both answers are correct. One solution is to develop a scale which awards points based on the amount of correct details in the answer. This solution requires test developers to prepare a list of all propositions in the passage. The other solution is to re-word the question: "**List at least two economic measures advocated by the speaker.**" This wording tells examinees how much information is expected in their response.

Yet another problem in scoring open-ended questions is presented by partially correct answers. One possible solution is to ask several highly proficient listeners to independently answer the questions, and to compile a list of their answers. The list is then given to the scorers to reduce the number of decisions they have to make. This may still leave the scorers with a small number of "far out" answers which will need arbitration.

Recall protocols

Recall protocols are normally administered in the following way; (1) a brief listening passage is recorded at normal speed; (2) a list is prepared of all facts or propositions contained in the passage; (3) students listen to the passage; (4) they are asked to write down everything they remember from the passage. More points may be awarded for recall of higher-level propositions than for details (Bernhard and James, 1985). Critics of this technique argue that it confounds listening comprehension with memory ability. In addition, recall protocols rely on writing on writing— a skill which may be even less developed than listening. Examinees may be reluctant to write down what they have understood if they are unsure of the grammar and spelling. The solution is for students to write the protocols in their native language. Finally, scoring of recall protocols is labor-intensive and requires training to ensure inter-rater reliability.

Non-verbal responses

Language teachers like to argue about the use of L1 in the classroom, and this argument spills over into discussions of testing procedures. Purists insist that L1 should be avoided at all costs, while pragmatists maintain a "whatever works best" position. From a psychometric perspective, the language of response is a source of measurement error because we cannot determine how much of the variation in the scores is attributable to listening comprehension, and how much to writing or speaking ability. Examinees may have understood a passage but were unable to demonstrate their comprehension through speaking or writing in L2. For this reason, at lower levels of proficiency, non-verbal responses are especially useful. A few examples of such responses are given below:

Test-taker hears:	A description of a house, a person, or weather.
Test taker sees:	Pictures of four different houses, persons, or weather scenes.
Task:	Circle the picture that corresponds to the description.
Test-taker hears:	A narrative about a specific event.
Test taker sees:	Pictures representing scenes from the narrative.
Task:	Place pictures in chronological order, based on the narrative.
Test-taker hears:	A narrative with a clear story line.
Test-taker sees:	Pictures of four possible outcomes of the story.
Task:	Select outcome most consistent with the story.

Test-taker hears:	A lecture on demographics.
Test-taker sees:	Graphs of charts representing different population trends.
Task:	Select graph or chart representing information in the lecture.

Test-taker hears:	Directions how to get somewhere.
Test-taker sees:	A city map.
Task:	Draw a line to indicate the route described in the directions.

PRESENTATION EFFECTS

Presentation effects have the potential of confounding listening comprehension with understanding instructions and test questions, as well as with differences in test administration. Some of the most obvious and controllable sources of error are described below.

Advance organizers

Listening in the real world normally occurs in context which helps listeners eliminate potentially ambiguous interpretations of the message, and to infer the meaning of unclearly heard or unfamiliar words or phrases. In addition, listeners normally have a purpose for listening in mind. This helps them decide what to concentrate on, and how to listen. In an effort to duplicate these conditions in test situations, it is common practice to give test-takers prelistening questions (Bacon, 1991). Lund (1991b) reported that listeners who were told to understand as much as they could and then write a recall protocol recalled fewer main ideas, fewer details, and produced more inappropriate interpretations of the text than listeners who were told what to focus on before they listened to a passage. Lund believes that unfocused instructions gave listeners little help in determining what to concentrate on, so that they tried to process everything indiscriminately. Respondents in an introspective study by Buck (1991) reported that question preview influenced their listening strategies, and made listening easier for them. However, Buck suggested that the effect of prelistening questions may, in fact, depend on the passage. Such questions may be helpful when listening to expository passages, crammed with facts, but not when listening to interesting stories with a clear story line. Note, however, that there are no empirical studies comparing the effects on listening comprehension of questions before listening with questions after listening.

Language of instructions and language of questions

The potential for reduced reliability of a listening test is even greater when it comes to presenting instructions and test questions in L2, especially in the case of lower-proficiency examinees, since it is impossible to determine how much of the variation in their listening scores can be attributed to their L2 listening ability and how much to their L2 reading comprehension. Whether you decide to present instructions and questions in L1 or L2, keep the wording short and simple, since your purpose is to test listening comprehension, not reading ability. If you decide to present questions in L2, keep in mind that it is difficult to simplify the language of multiple-choice items. It is also a good idea to offer a sample passage for practice to ensure that test-takers understand what is expected of them. A sample question provides a warmup for students who may otherwise miss answering the first test question while trying to adjust to the format of the test.

Uniformity of presentation

You should make sure that you standardize the way you administer your listening test. If you present a listening passage live to several classes, it will not be possible for you to account for variations in speed, loudness, emphases, pauses, acoustics, and background noise. If your test is administered by different instructors, there will also be no way to account for the potential impact of the difference in their voices. Therefore, it is essential that you record the passages you want to include in your test.

In addition, you should keep constant the number of times the passages are repeated, as well as time to complete responses. You should keep in mind that repeated presentations of a listening passage will not be particularly helpful to low-level listeners, whereas advanced listeners will be more likely to profit from hearing the passage several times (Lund, 1991a). In any case, the number of repetitions should be kept constant from one test administration to another.

It is also essential that you give exactly the same instructions to all groups of test-takers. For instance, if one group is warned that there is a penalty for guessing, and another group is not, examinees in the two groups will adopt different test-taking strategies and that, in turn, will affect their test performance.

REFINING YOUR TEST

Chances are that the first time you use a new test, some of the items will turn out to be unreliable. A few relatively simple steps can go a long way towards increasing the reliability of your test without doing complicated and time-consuming statistical analyses. First, give the test to a few people without having them listen to the passages to find out if they can correctly answer any of the questions without the benefit of having heard the passages. If they can answer some of the questions correctly, it means that they are based on extratextual information and can be answered solely on the basis of familiarity with the topic, logical reasoning, and other types of extralinguistic knowledge. These items should be discarded.

Secondly, pilot the test in one of your classes, and analyze the results. Specifically, look for good and bad items. In norm-referenced tests, a good item is one of average difficulty, i.e., one which is answered correctly by about fifty percent of the examinees. In addition, a good item is one which correlates well with the total scores, that

is, it ranks test-takers in approximately the same way as the total test scores. Items that were answered incorrectly by test-takers who generally did well on the test, and items that were answered correctly by those who did poorly on the test as a whole should be discarded or re-worded. In addition, items that were answered correctly or incorrectly by most examinees are non-discriminating and, they too should be discarded. If you repeat this procedure several times, you will end up with a test that is reliable enough for purposes of formative evaluation. However, if more important decisions ride on the results of the test, you should consider adopting a standardized test, or seek the help of a psychometrician.

CONCLUSION

In this paper, I have made some practical suggestions on how to make classroom tests of listening comprehension more valid and reliable through careful selection of listening passages, and creation of listening tasks that reflect cognitive operations involved in real life listening. These suggestions must be construed as tentative pending the development of a more fully elaborated model of listening comprehension.

REFERENCES

Aly, Anwar Amer. 1993. "Teaching EFL Students to Use a Test-Taking Strategy." *Language Testing* 10(1): 72-77.

Bacon, Susan M. 1991. "Assessing Foreign Language Listening: Processes, Strategies, and Comprehension: in Richard V. Teschner (ed): *Assessing Foreign Language Proficiency of Undergraduates*, pp. 105-20. Boston, Mass.: Heinle and Heinle.

Berne, Jane Ellen. 1992. "The Effects of Text Type, Assessment Task, and Target Language Experience on Foreign Language Learners' Performance on Listening Comprehension."Unpublished Doctoral Dissertation. University of Illinois at Urbana-Champaign.

Bernhard, Elizabeth B., and Charles J. James. 1987. "The Teaching and Testing of Comprehension in Foreign Language Learning" in Diane W. Birckbichler (ed): *Proficiency, Policy, and Professionalism in Foreign Language Education*, pp. 63-81. Lincolnwood, Ill.: National Textook Company.

Blau, Eileen K. 1990 "The Effect of Syntax, Speed and Pauses on Listening Comprehension." *TESOL Quarterly* 24(4):746-53.

Blau, Eileen K. 1991. "The Effect of Pauses and Hesitation Markers on Listening Comprehension." Unpublished paper.

Buck, Gary. 1991. "The Testing of Listening Comprehension: An Introspective Study." *Language Testing* 8(1): 67-91.

Carterette, E.C., and M. Hubbard Jones. 1974. *Informal Speech*. Berkeley, Calif.: University of California Press.

Chafe, Wallace L. 1985. "Linguistic Differences Produced by Differences between Speaking and Writing" in David Olson, Nancy Torrance, and Angela Hildyard (eds): *Literary Language and Learning*, pp. 105-23. Cambridge: Cambridge University Press.

Chaudron, Craig. 1983. "Simplification of Input: Topic Reinstatements and Their Effects on L2 Learners' Recognition and Recall." *TESOL Quarterly* 17(3): 437-58.

Chaudron, Craig, and Jack C. Richards. 1986. "The Effect of Discourse Markers on the Comprehension of Lectures." *Applied Linguistics* 7(2):113-27.

Chiang, Chung Shing, and Patricia Dunkel. 1992. "The Effect of Speech Modification, Prior Knowledge, and Listening Proficiency on EFL Lecture Learning." *TESOL Quarterly* 26(2):345-74.

Derwing, Tracy. 1989. "Information Type and Its Relation to Nonnative Speaker Comprehension." *Language Learning* 39(2):157-72.

Educational Testing Service (ETS). 1990. *Comprehensive Russian Proficiency Test*. Princeton, NJ: Educational Testing Service.

Glisan, Eileen W. 1985. "The Effect of Word Order on Listening Comprehension and Pattern Retention: An Experiment in Spanish as a Foreign Language." *Language Learning* 35(3):443-69.

Griffith, Roger. 1990. "Speech Rate and Nonnative Speaker Comprehension: A Preliminary Study in the Time-Benefit Analysis." *Language Learning* 40(3):311-36.

Griffith, Roger. 1992. "Speech Rate and Listening Comprehension: Further Evidence of the Relationship." *TESOL Quarterly* 26(2):385-91.

Henrichsen, Lynn E. 1990. "Sandhi-Variation: A Filter of Input for Learners of ESL." *Language Learning* 34(3):103-26.

Hieke, A.E. 1987. "The Resolution of Dynamic Speech in L2 Listening." *Language Learning* 37(1): 123-40.

Hildyard, Angela, and David R. Olson. 1982. "On the Comprehension and Memory for Oral vs. Written Discourse" in Deborah Tannen (ed): *Spoken and Written Language: Exploring Orality and Literacy*, pp. 19-33. Norwood, NJ: Ablex.

Hron, Aemilian, Ingeborg Kurbjuhn, Heinz Mandl, and Wolfgang Schnotz. 1985. "Structural Inferences in Reading and Listening." *Inferences in Text Processing*, pp. 221-45. North-Holland: Elsevier Science Publishers.

Jacobs, George, Wiladlak Chuawanlee, Bert K. Itoga Jr., Diane Sakumoto, Susan Saka, and Kenneth A. Meeham. 1988. "The Effect of Pausing on Listening Comprehension." Paper presented at The Eighth Second Language Research Forum. Honolulu, Hawaii, March 3-6.

Joiner, Elizabeth G. 1990. "Choosing and Using Videotexts." *Foreign Language Annals* 23(1):53-64.

Kelch, Ken. 1985. "Modified Input as an Aid to Comprehension." *Studies in Second Language Acquisition* 7(1):81-89.

King, Paul E., and Ralph R. Behnke. 1989. "The Effect of Time-Compressed Speech on Comprehensive, Interpretive, and Short-Term Listening." *Human Communication Research* 15(2): 428-41.

Laviosa, Flavia. 1991. "An Investigation of the Listening Strategies of Advanced Learners of Italian as a Second Language." Paper presented at Bridging Theory and Practice in the Foreign Language Classroom Symposium. Loyola College, Maryland, October 18-20.

Long, Donna Resigh. 1990. "What You Don't Know Can't Help you: An Exploratory Study of Background Knowledge and Second Language Listening Comprehension." *Studies in Second Language Acquisition* 12(1):65-80.

Lund, Randall J. 1991a. "A Comparison of Second Language Listening and Reading Comprehension." *Modern Language Journal* 75(2):196-211.

Lund, Randall J. 1991b. "The Effects of Listening Tasks on Comprehension." Paper presented at the Research Perspectives in Adult Language Learning and Acquisition Annual Symposium. Columbus, Ohio, October 11-12.

Markham, Paul, and Michael Latham. 1987. "The Influence of Religion-Specific Background Knowledge on the Listening Comprehension of Adult Second-Language Learners." *Language Learning* 37(2):157-70.

Meyer, Bonnie, J.F., and George W. McConkie. 1973. "What is Recalled after Hearing a Passage?" *Journal of Educational Psychology* 65(1):109-17.

Mueller, Gunther. 1980. "Visual Contextual Cues and Listening Comprehension: An Experiment." *Foreign Language Annals* 64(3):335-40.

Phillips, June K. 1990. "An Analysis of Text in Video Newscasts: A Tool for Schemata Building in Listening." *Georgetown University Round Table on Language and Linguistics*. Washington, DC: Georgetown University Press.

Rubin, Joan. 1994. "A Review of Second Language Listening Comprehension Research." *The Modern Language Journal* 78(2):199-221.

Schmidt-Rinehart, Barbara. 1992. "The Effects of Topic Familiarity on the Listening Comprehension of University Students of Spanish." Unpublished Doctoral Dissertation. Ohio State University.

Shohamy, Elana, and Ofra Inbar. 1991. "Validation of Listening Comprehension Tests: The Effect of Text and Question Type." *Language Testing* 8(1):23-50.

Tannen, Deborah. 1982. "The Oral/Literate Continuum in Discourse" in Deborah Tannen (ed): *Spoken and Written Language: Exploring Orality and Literacy*, pp. 1-16. Norwood, NJ: Ablex.

Tannen Deborah. 1985. "Relative Focus of Involvement in Oral and Written Discourse" in David R. Olson, Nancy Torrance, and Angela Hildyard (eds): *Literary Language and Learning*, pp. 124-47. Cambridge: Cambridge University Press.

Thompson, Irene. 1993. "An Investigation of the Effects of Texts and Tasks on Listening Comprehension: Some Evidence from Russian." *Georgetown University Round Table on Languages and Linguistics*, pp. 294-305. Washington, DC: Georgetown University Press.

Thompson, Irene, and Joan Rubin. "Can Strategy Training Improve Listening Comprehension?" (Forthcoming).

Wolvin, Andrew D., and Carolyn Gwynn Coakley. 1985. *Listening* (2nd ed.) Dubuque, Iowa: William C. Brown Company.